Logistic Regression

* Binary classification and have the probability of each class

1. we need to limit $w^T x + b$ to $[0,1]$, so use sigmoid function ← good for binary classification
   ↳ probability

$\hat{y} = \theta(w^T x + b)$ $\quad \theta(z) = \frac{1}{1+e^{-z}}$ $\quad$ for classification: $\hat{y} = \begin{cases} 1 & \text{if } \theta(z) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}$

2. when you add the sigmoid activation function, you have to make sure that the loss function is convex

loss

the cross-entropy loss function is common. why? It is derived from maximum likelihood estimation.

for 1 observation, we are given features x_i, weights θ, and value y_i $\quad y \in \{0,1\}$

$$\theta(x_i, \theta) = \frac{1}{1+e^{-(\theta_0 + \theta^T x_i)}} \quad \text{Sigmoid}$$

probability of y_i given x_i and θ

$P(y=1 \mid x_i, \theta) = \theta(x_i, \theta)$
$P(y=0 \mid x_i, \theta) = 1 - \theta(x_i, \theta)$] recall that there are only 2 classes

$$P(y \mid x_i, \theta) = [\theta(x_i, \theta)]^{y_i} [1 - \theta(x_i, \theta)]^{1-y_i}$$

for n observations, it is the product of the probability.
we want the weights θ that maximize the probability

$$\mathcal{L}(x_i, \theta) = \max_\theta \prod_{i=1}^n [\theta(x_i, \theta)]^{y_i} [1 - \theta(x_i, \theta)]^{1-y_i}$$

$\hat{y} = \theta(x_i, \theta)$

$$= \max_\theta \prod_{i=1}^n [\hat{y}_i]^{y_i} [1 - \hat{y}_i]^{1-y_i}$$ ↓ take log to turn products to sum

this is called log loss $\quad \ln(\mathcal{L}(x_i, \theta)) = \max_\theta \sum_{i=1}^n \ln([\hat{y}_i]^{y_i} [1-\hat{y}_i]^{1-y_i})$

$$f(\theta) = \max_\theta \sum_{i=1}^n y_i \ln(\hat{y}_i) + (1-y_i)\ln(1-\hat{y}_i)$$ ← the (y_i) and $(1-y_i)$ terms are like an indicator variable $\mathbb{1}\{y_i = k\}$ $\quad y \in \{0,1\}$

↳ or cross entropy loss function (for binary case) $\quad y \in \{0,1\}$

Multi-class Classification

one versus all
- K binary classifiers that say it belongs to the class or not. return model with highest probability

all versus all
-

Softmax Regression

* essentially a neural network

- extends logistical regression to multiple classes and gives likelihoods of each label

in logistical regression, since there are only (2) classes, we did:
$P(y=1 \mid x_i, \theta) = a$ and $P(y=0 \mid x_i, \theta) = 1 - a$

Softmax regression when $K=2$ is logistical regression

- each class out of K classes has its own weight vector w_K
- score for class K using weight vector w_K: $w_K^T x$

probability for class K:

$$P(y=K \mid x, w_K) = \frac{e^{w_K^T x}}{\sum_{j=1}^K e^{w_j^T x}}$$

softmax:

$$f(x, w) = \frac{1}{\sum_{j=1}^K e^{w_j^T x}} \begin{bmatrix} e^{w_1^T x} \\ \vdots \\ e^{w_K^T x} \end{bmatrix}$$ where $W = \begin{bmatrix} — w_1^T — \\ — w_2^T — \\ \vdots \\ — w_K^T — \end{bmatrix}$

Sigmoid gradient $\nabla_z g(z) = g(z)[1-g(z)]$

$g(z_i) = \frac{1}{1+e^{-z_i}}$

$$\begin{aligned}
\nabla_z g(z) &= \nabla_z \left[\frac{1}{1+e^{-z}}\right] \\
&= \nabla_z [1+e^{-z}]^{-1} \\
&= -[1+e^{-z}]^{-2} \cdot \nabla_z [1+e^{-z}] \\
&= -[1+e^{-z}]^{-2} \cdot (e^{-z}) \cdot \nabla_z (-z) \\
&= (1+e^{-z})^{-2} e^{-z} \\
&= \frac{e^{-z}}{(1+e^{-z})^2} \\
&= \frac{1+e^{-z}-1}{(1+e^{-z})^2} \\
&= \frac{1+e^{-z}}{(1+e^{-z})^2} - \frac{1}{(1+e^{-z})^2} \\
&= \frac{1}{1+e^{-z}} - \frac{1}{(1+e^{-z})^2} \\
&= \frac{1}{1+e^{-z}}\left(1 - \frac{1}{1+e^{-z}}\right) \\
&= g(z_i)(1-g(z_i))
\end{aligned}$$

recall $g(z_i) = \frac{1}{1+e^{-z}}$

Logistic Regression Gradient Descent (lecture 12 notes)

$$\ell(\theta) = -\sum_{i=1}^{n} [y_i \ln(h_\theta(x_i)) + (1-y_i)\ln(1-h_\theta(x_i))]$$

we want $\theta = \theta - \eta \nabla_\theta \ell(\theta)$ for gradient descent

$$\begin{aligned}
\ell(\theta) &= -\sum_{i=1}^{n} [y_i \ln(h_\theta(x_i)) - y_i \ln(1-h_\theta(x_i)) + \ln(1-h_\theta(x_i))] \\
&= -\sum_{i=1}^{n} \left[y_i \ln\left(\frac{h_\theta(x_i)}{1-h_\theta(x_i)}\right) + \ln(1-h_\theta(x_i))\right]
\end{aligned}$$

$\frac{h_\theta(x_i)}{1-h_\theta(x_i)} = \frac{\left(\frac{1}{1+e^{-(\theta_0+\theta^T x_i)}}\right)}{\left(\frac{e^{-(\theta_0+\theta^T x_i)}}{1+e^{-(\theta_0+\theta^T x_i)}}\right)} = \frac{1}{e^{-(\theta_0+\theta^T x_i)}} = e^{\theta_0+\theta^T x_i}$

$h_\theta(x_i) = \frac{1}{1+e^{-(\theta_0+\theta^T x_i)}}$

$1-h_\theta(x_i) = \frac{1+e^{-(\theta_0+\theta^T x_i)}}{1+e^{-(\theta_0+\theta^T x_i)}} - \frac{1}{1+e^{-(\theta_0+\theta^T x_i)}} = \frac{e^{-(\theta_0+\theta^T x_i)}}{1+e^{-(\theta_0+\theta^T x_i)}} = \frac{e^{-(\theta_0+\theta^T x_i)} \cdot e^{\theta_0+\theta^T x_i}}{(1+e^{-(\theta_0+\theta^T x_i)}) \cdot e^{\theta_0+\theta^T x_i}} = \frac{1}{e^{\theta_0+\theta^T x_i}+1}$

$$= -\sum_{i=1}^{n} \left[y_i \ln(e^{\theta_0+\theta^T x_i}) + \ln\left(\frac{1}{1+e^{\theta_0+\theta^T x_i}}\right)\right]$$

$$\ell(\theta) = -\sum_{i=1}^{n} [y_i(\theta_0+\theta^T x_i) - \ln(1+e^{\theta_0+\theta^T x_i})]$$

$$\begin{aligned}
\nabla_\theta \ell(\theta) &= -\sum_{i=1}^{n} [y_i \cdot \nabla_\theta(\theta_0+\theta^T x_i) - \nabla_\theta \ln(1+e^{\theta_0+\theta^T x_i})] \\
&= -\sum_{i=1}^{n} [y_i \cdot x_i - \left(\frac{1}{1+e^{(\theta_0+\theta^T x_i)}}\right) \cdot \nabla_\theta (e^{\theta_0+\theta^T x_i}+1)] \\
&= -\sum_{i=1}^{n} [y_i \cdot x_i - \left(\frac{1}{1+e^{(\theta_0+\theta^T x_i)}}\right) \cdot (e^{\theta_0+\theta^T x_i}) \cdot \nabla_\theta (\theta_0+\theta^T x_i)] \\
&= -\sum_{i=1}^{n} [y_i \cdot x_i - x_i \cdot \left(\frac{e^{(\theta_0+\theta^T x_i)}}{1+e^{(\theta_0+\theta^T x_i)}}\right)] \\
&= -\sum_{i=1}^{n} [y_i \cdot x_i - x_i \cdot \left(\frac{1}{1+e^{-(\theta_0+\theta^T x_i)}}\right)] \\
&= -\sum_{i=1}^{n} [y_i \cdot x_i - x_i \cdot h_\theta(x_i)]
\end{aligned}$$

$\hat{y} = h_\theta(x_i) = \frac{1}{1+e^{-(\theta_0+\theta^T x_i)}}$

$$= -\sum_{i=1}^{n} [y_i \cdot x_i - x_i \cdot \hat{y}_i]$$

$$\nabla_\theta \ell(\theta) = \sum_{i=1}^{n} x_i (\hat{y}_i - y_i)$$

$$\theta_j = \theta_j - \eta \left(\sum_{i=1}^{n} [\hat{y}_i - y_i] x_{ij}\right)$$

multiply by $\frac{1}{n}$ to average over the loss

$\nabla_\theta \ell(\theta) = -\sum_{i=1}^{n} (\hat{y}_i - y_i) x_{ij}$

$\hat{y} = h_\theta(x_i)$ in terms of weight vector θ and the input x_i

recall $\nabla_\theta \ell(\theta) = \begin{bmatrix} \frac{\partial}{\partial \theta_1}[\ell(\theta)] \\ \frac{\partial}{\partial \theta_2}[\ell(\theta)] \\ \vdots \\ \frac{\partial}{\partial \theta_0}[\ell(\theta)] \end{bmatrix}$

$\theta_0 + \theta^T x_i$

$\theta = \begin{bmatrix} \theta_1 \\ \theta_2 \\ \vdots \\ \theta_0 \end{bmatrix} \quad x_i = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_0 \end{bmatrix}$

$= \theta_0 + \theta_1 x_1 + \ldots + \theta_0 x_0$

$\nabla_\theta [\theta_0 + \theta^T x_i] = x_i$